Dr. (Amos) Xinshao Wang: Data + Modelling

Vision and Endeavours: To Design and Build Data/Label/Time-Efficient, Robustly Reliable and Transparent AI for Diverse Applications (Omics/Biosequences, Healthcare, NLP, CV, etc).

Google Scholar: yOBhB7UAAAAJ
Github: https://github.com/XinshaoAmosWang
Homepage+Blogs: https://xinshaoamoswang.github.io/

Email: xinshaowang@gmail.com
Phone: +44 (0) 7712 114316

2023-04-19 Wed, Invited Talk@Synteny

Outline

Outline

Self introduction: Overview

2013

China National Scholarship ×3

APSIPA 2015 Oral

Intern at YouTu Lab, Tencent

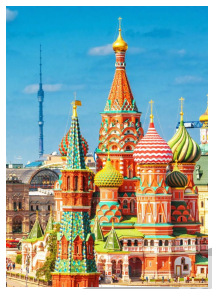
2017

* B.Eng. at NWAFU ended
* Master offer from PKU
* PhD offer from CUHK
* PhD offer from QUB and Anyvision

Hawaii

aaai

CVPR LONG BEACH CALIFORNIA June 16-20, 2019



MACHINE LEARNING SUMMER SCHOOL

* PhD study at QUB ended
* Postdoc offer from Oxford University

* Papers: AAAI Oral, 2 CVPRs, TPAMI

* Reviewers: CVPR, ICCV, AAAI, NeurIPS, TPAMI, TNNLS, Knowledge Based Systems

2020-06

AI for health care

2020-12

* A senior ML Research Scientist at Zenith Ai, UK

* A founding technical member at Zenith Ai

* Acquired by Opentrons, a Unicorn

* Working on Omics AI

* Visit Scholar at Oxford University

* Papers:
  - TSRML Workshop of NeurIPS, 2022,
  - NeurIPS, 2022

Global Talent visa, by Royal Academy of Engineering

B.Eng. | PhD | PostDoc | Industrial Research

Outline

Work and experience

1. Deep distance metric learning
   - CVPR 2019 and TPAMI 2022
   - AAAI 2019 Oral
2. Robust deep learning, model calibration and uncertainties:
   - CVPR 2021
   - Co-supervising a PhD student: Trustworthy and Socially Responsible ML Workshop, NeurIPS 2022
   - Trustworthy and Reliable ML Workshop 2023, ICLR "IMAE for Noise-Robust Learning: Mean Absolute Error Does Not Treat Examples Equally and Gradient Magnitude's Variance Matters" (65 citations)
3. (Oxford) Postdoc, Visit Scholar on AI health care (e.g., ECG)
4. (Zenith Ai) Sr. Researcher on Omics AI (e.g., DNA, tRNA, protein, amino acid, ribosome, structure-sequence-function)

Technical novelty and principle

Example weighting is universal in deep learning

We define our interpretation of example weighting [6]:

Definition (Example Weighting). *In gradient-based optimisation, the derivative of an example can be interpreted as its effect on the update of a model. Therefore, a derivative's magnitude function equals to a weighting scheme.*

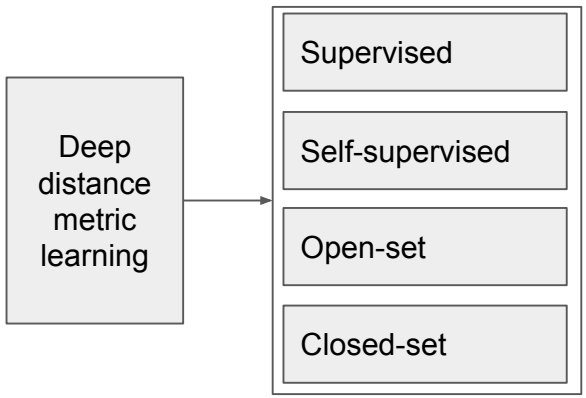
Accordingly, a change of the derivative magnitude function, is implicitly equivalent to, modifying an example weighting scheme.

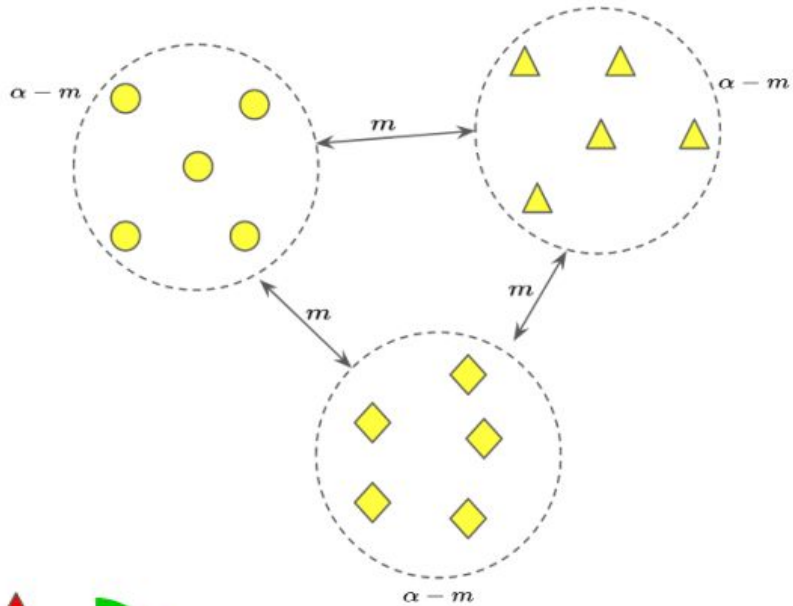
Intuitive research motivations:
(1) Not all training examples are created equal!
(2) Sampling matters!

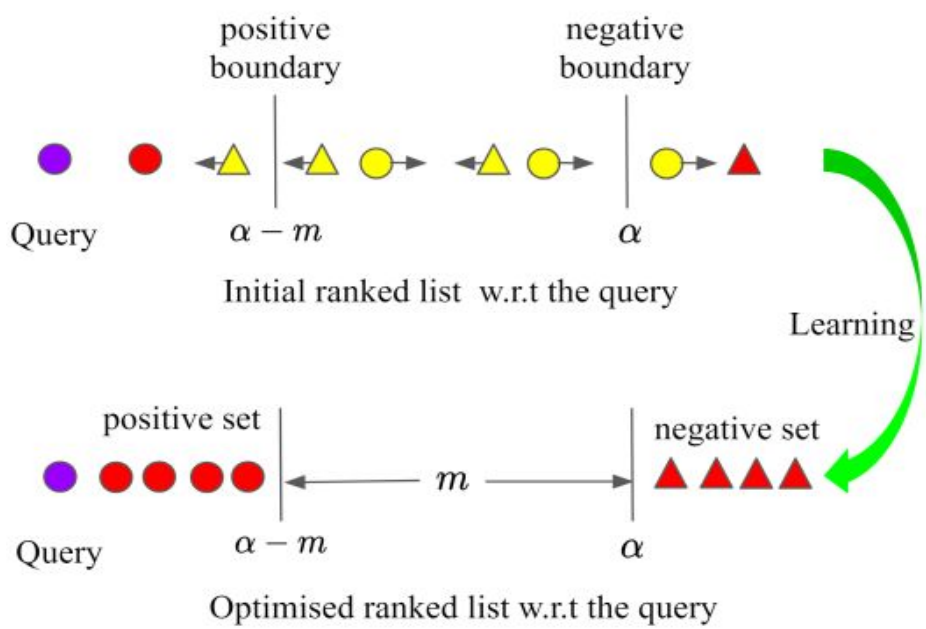
Outline

Structured learning on selected data

CVPR 2019 and TPAMI 2022 [8]





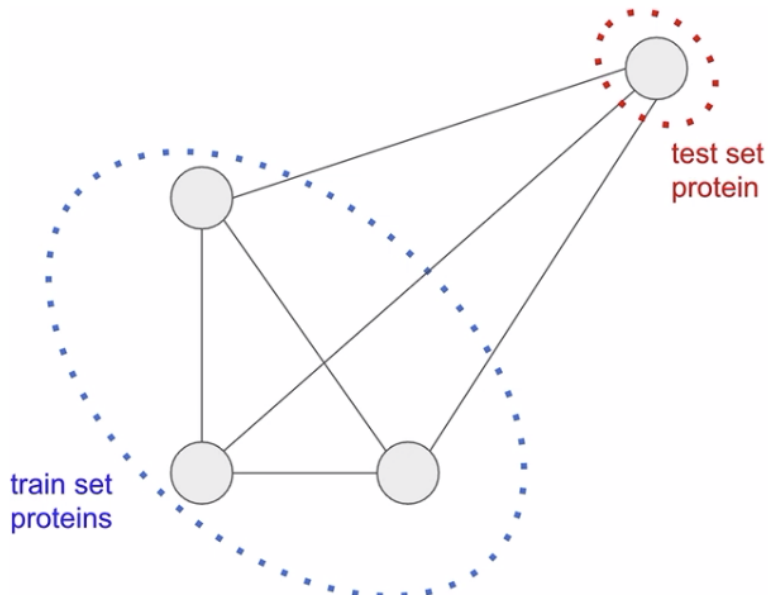


Outline

Learning objectives of robust deep learning

What is the meaning of robustness here?

1. **To learn meaningful patterns** on semantically clean data.
2. **Without fitting errors/bias**.
3. **Generalisation** to **unseen** data.



Build training and testing datasets properly [1].
How about the validation dataset?

Outline

Contribution: disclose the learning dynamics [7]

DNN has strong fitting capability, *but we find*:

Deep models easily fit random noise.

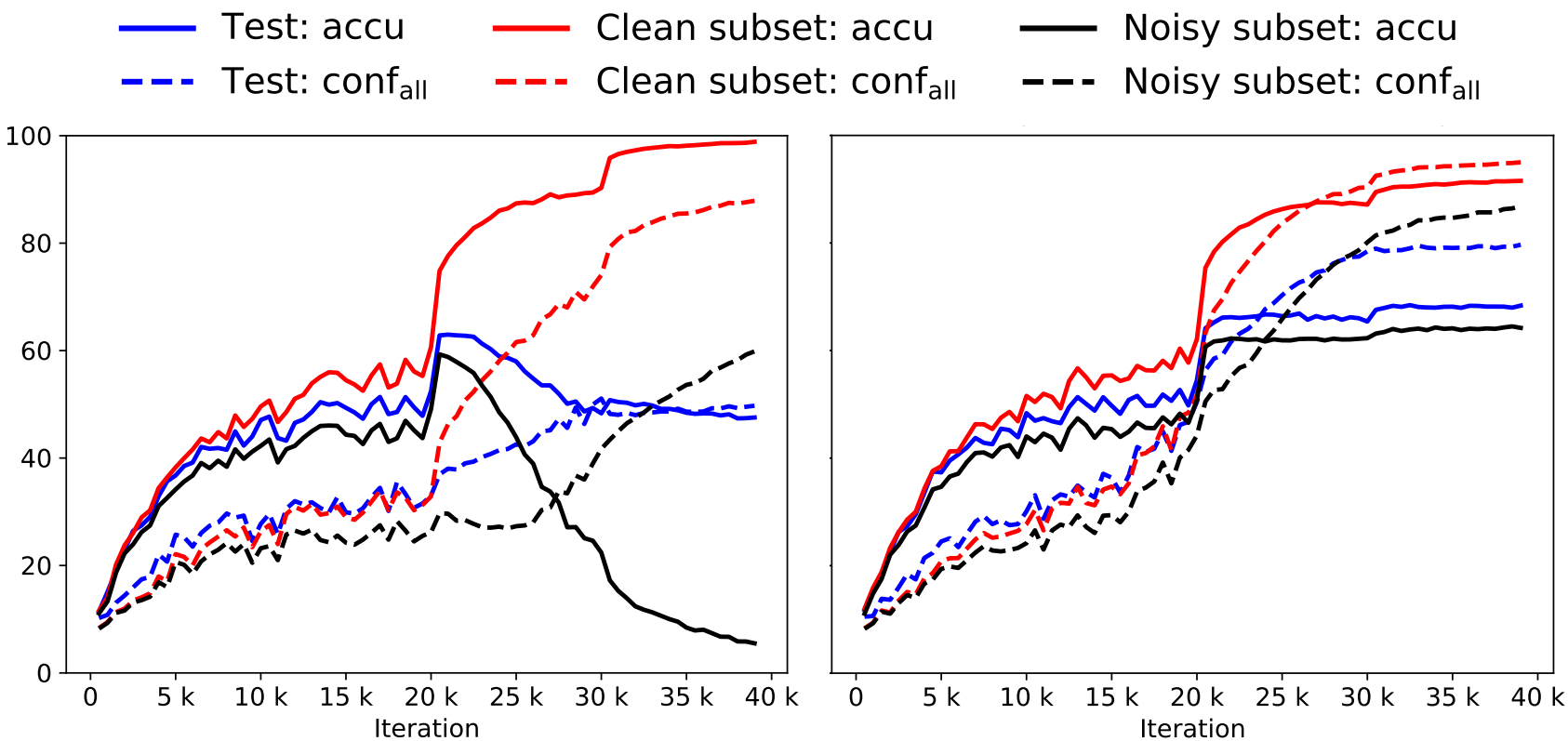
Deep networks learn simple semantic patterns before fitting noise.

Modern deep neural works tend to be over-confident.

(Ours: miscalibration under the noise) Deep neural networks become less confident of learning semantic patterns before fitting noise when the label noise rises.

Contribution: propose the ProSelfLC to promote confident and accurate learning [7]

1. To reward a low-entropy status other than penalise.
2. To promote model calibration.



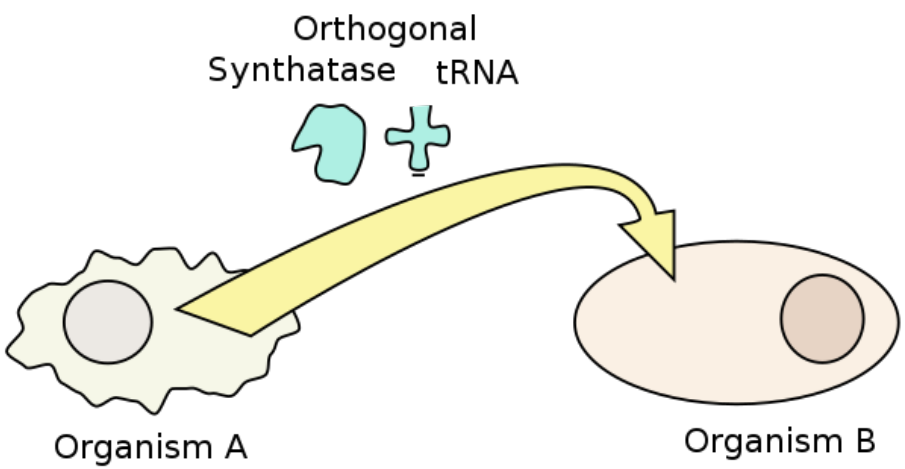
(a) Standard training. (b) ProSelfLC with an AT (ours).

Figure: accu and $\mathrm{conf}_{\mathrm{all}}$ when training ResNet18 on CIFAR-100. The noise rate is $r = 40\%$.

Outline

Cutting-edge OTS research: triplet design

Orthogonal (aaRS, tRNA, ncAA)



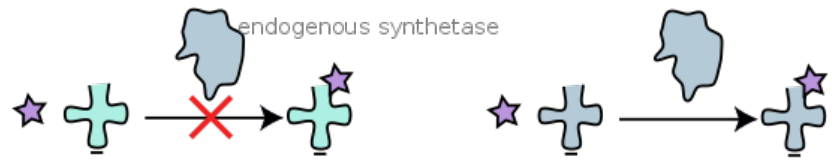
With the conditions that

I

Orthogonal synthetase can aminoacylate only the orthogonal tRNA

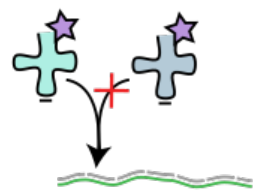
II

Endogenous synthases cannot aminoacylate the orthogonal tRNA



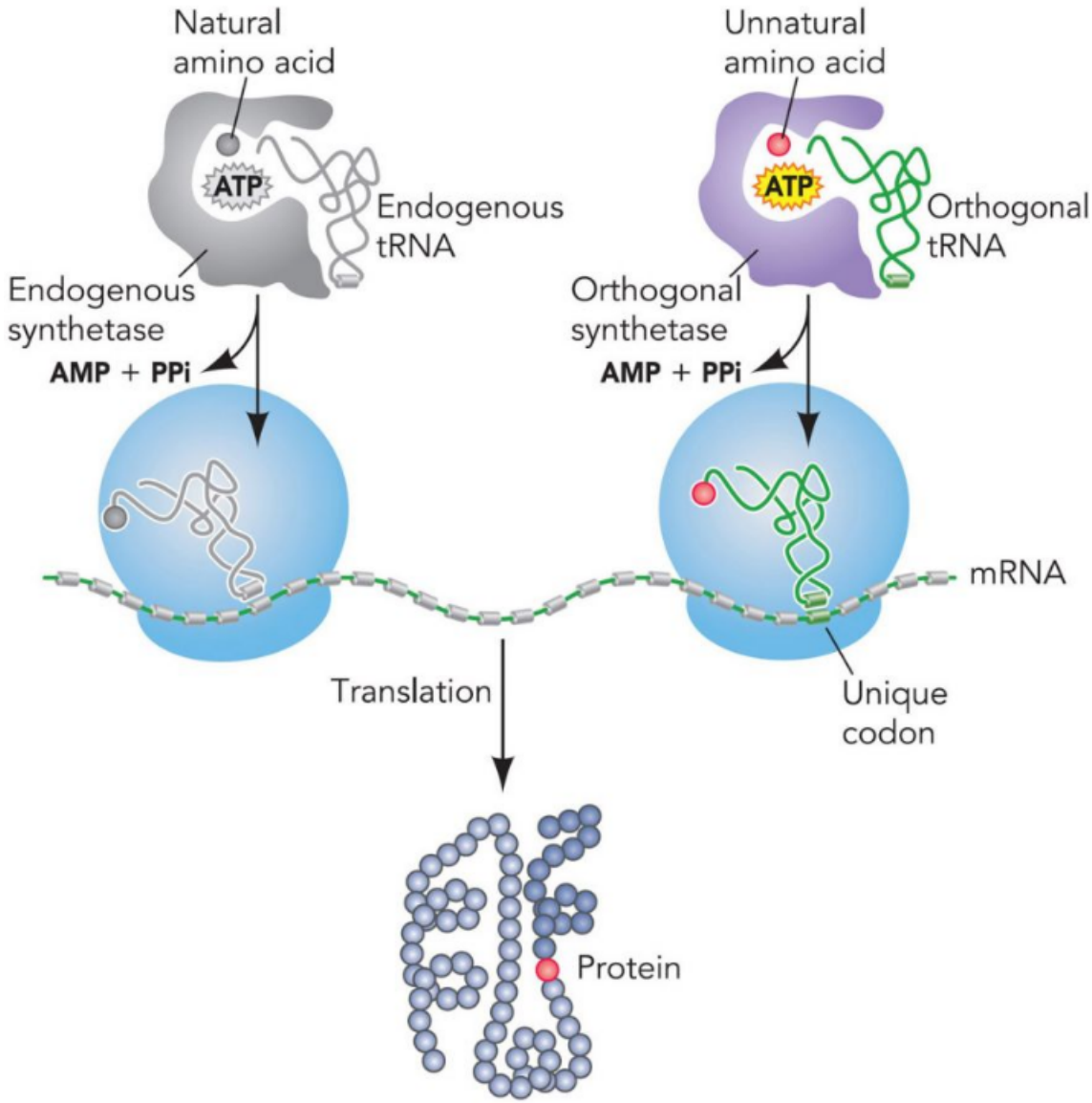
III

The orthogonal tRNA binds an unallocated codon



Cutting-edge OTS research: triplet design

Orthogonal (aaRS, tRNA, ncAA)

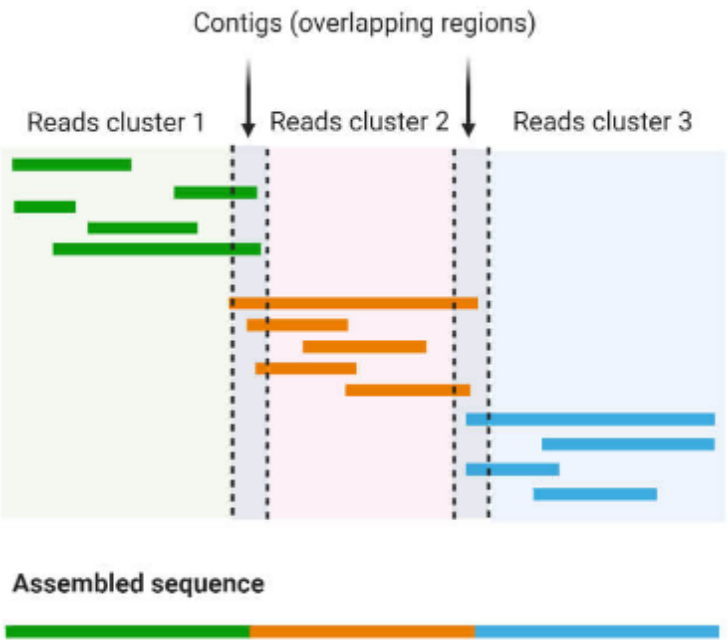


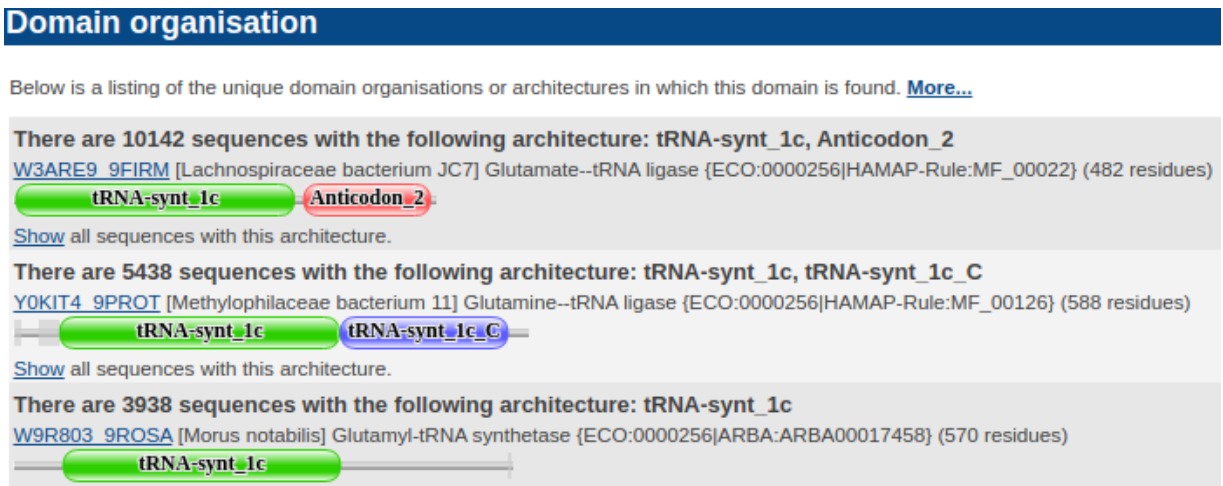
Outline

Bioinformatics

Sequence alignment and distance metrics

1. Gene synthesis: seq. optimisation + overlap-based assembly
2. NGS data analysis (e.g., variants calling + expression analysis)
   - Alignment: Bowtie2/minimap2 + Samtools
   - Reads counting
3. Alignment-based distance metrics
   - MMSeq2
   - Pfam domain database + HMMer-based distance

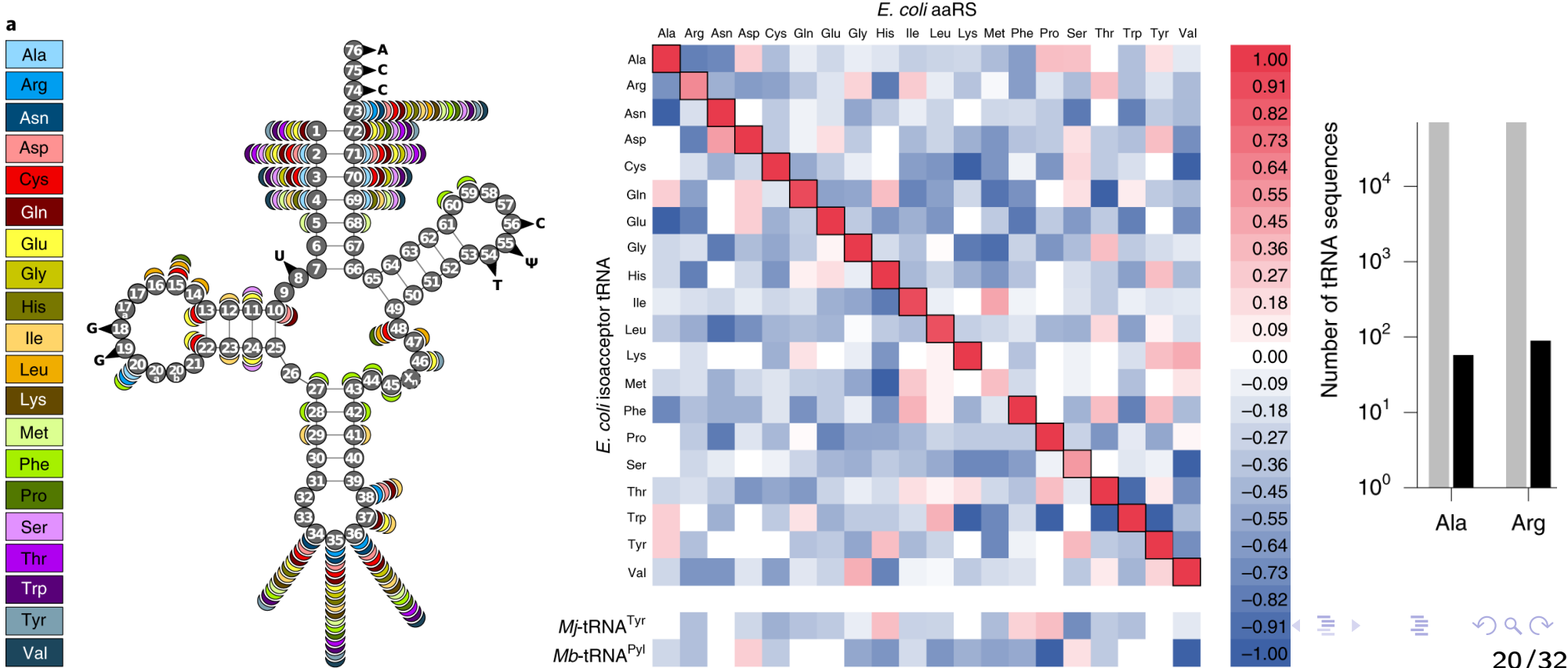




Bioinformatics for virtual screening

Revisiting Neil's question on sequences from diverse organisms

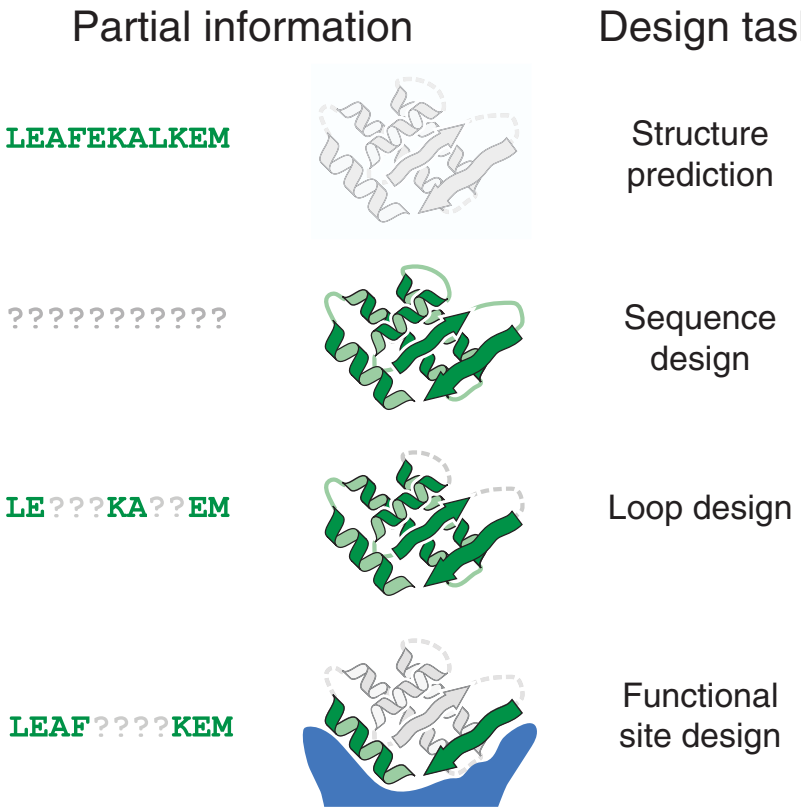
1. tRNA distance calculation [2]
   - Secondary structure and the canonical numbering scheme.
   - Identity elements, responsible for recognising cognate aaRS.
2. Annotating/validating via wetlab experiments: differential gene expression analysis.

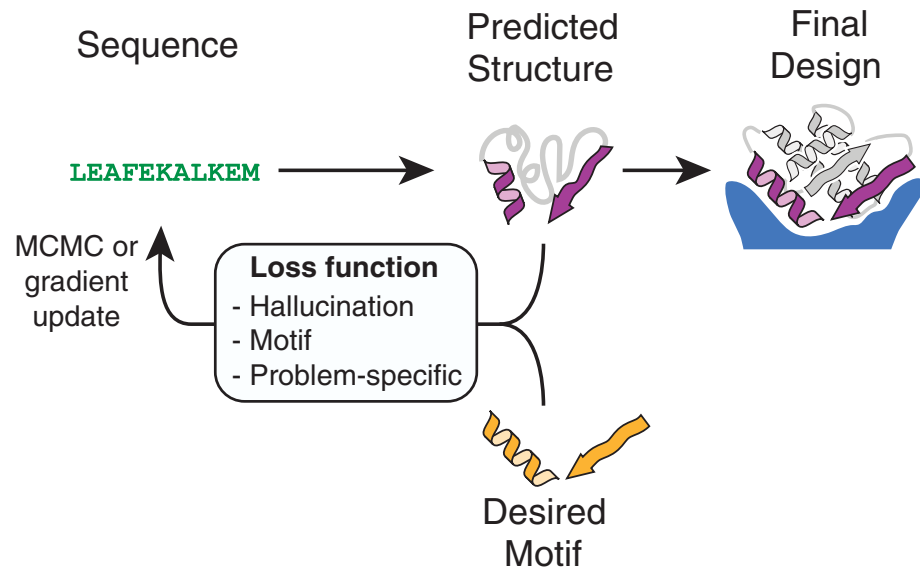


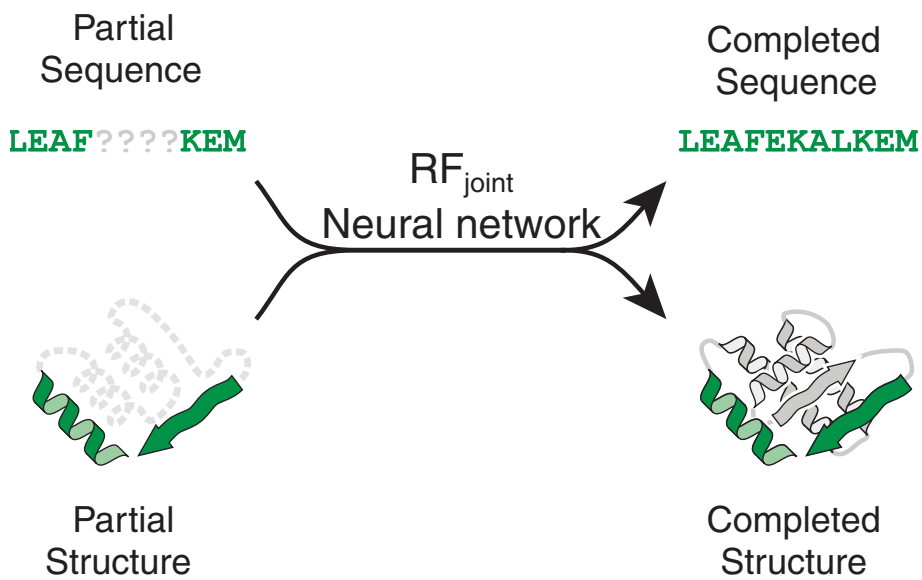
Outline

Not-solved ideas: hierarchical objectives

StructureMotif-Loop-Site; Sequence-Domain-Site



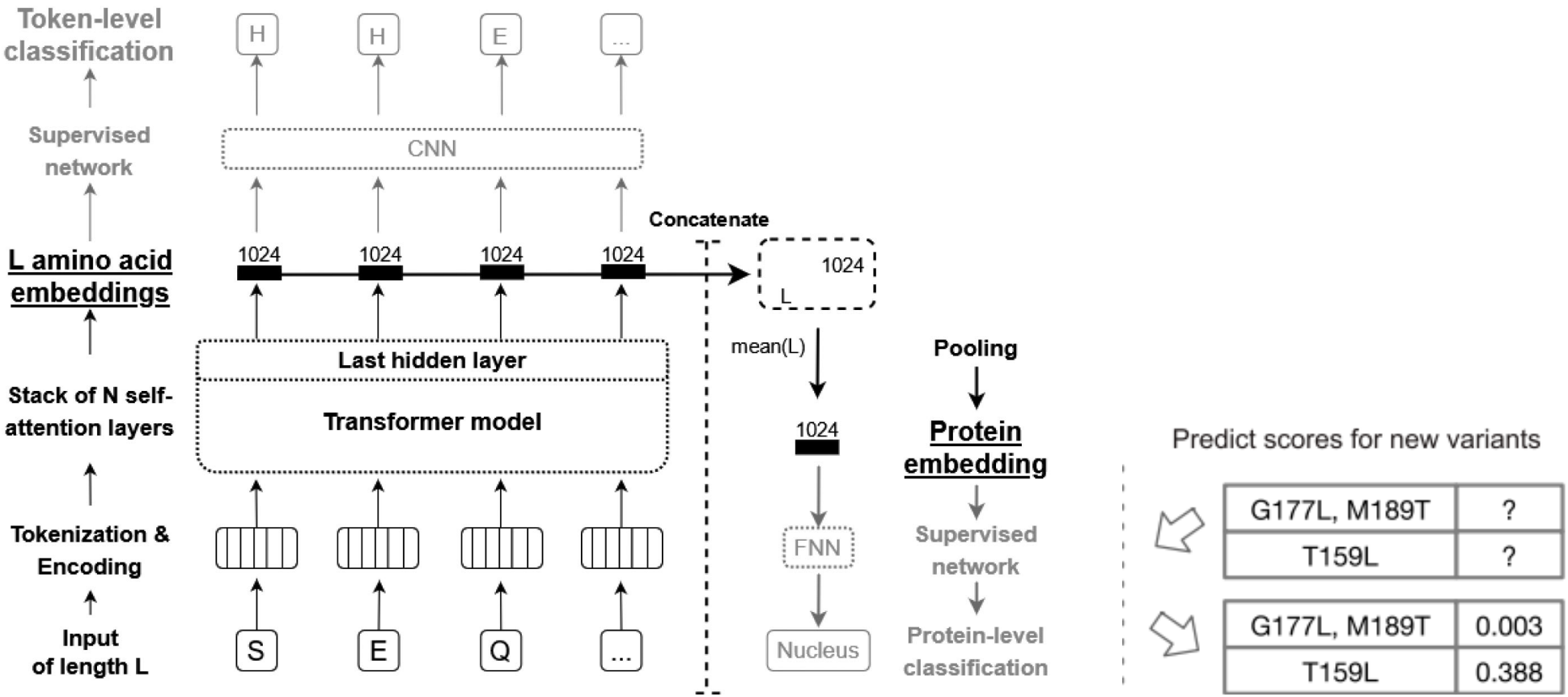




[5] "Scaffolding protein functional sites using deep learning" Science (2022).

Mutational effect: single and co-mutation

Active AI for selective/evolutionary screening



[3] "ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning." IEEE Transactions on Pattern Analysis and Machine Intelligence (2022).
[4] "Neural networks to learn protein sequence–function relationships from deep mutational scanning data." Proceedings of the National Academy of Sciences (2021).
[7] "ProSelfLC: Progressive Self Label Correction Towards A Low-Temperature Entropy State." Ours under peer review.

Public/Proprietary data collection, curation

Interpretative ML: backbone + functional loops/regions/sites

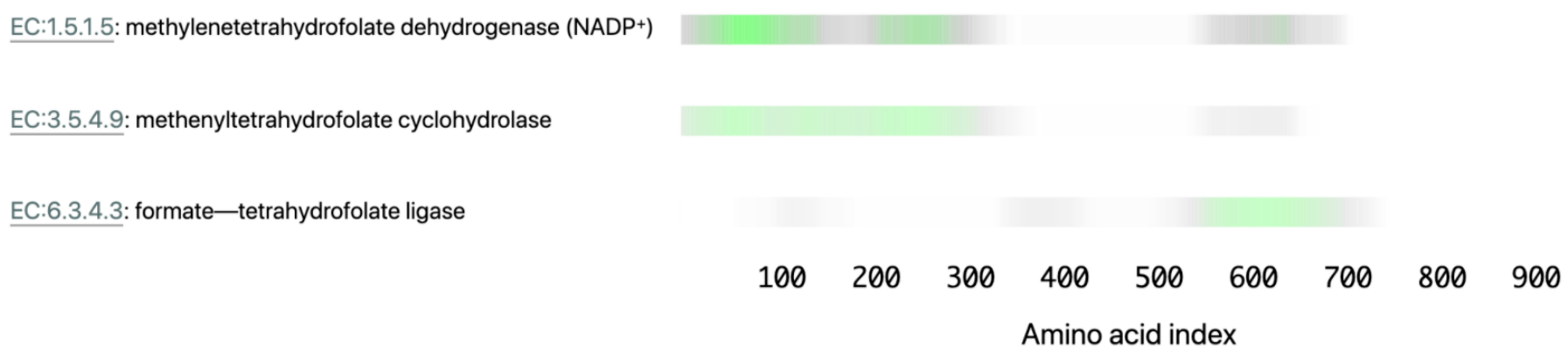
Domain visualisation of the same EC function:

UniProt's annotations versus model's predictions:

A

Position(s)	Description	Graphical view	Length
2 – 305	Methylenetetrahydrofolate dehydrogenase and cyclohydrolase		304
306 – 935	Formyltetrahydrofolate synthetase		630

B



Outline

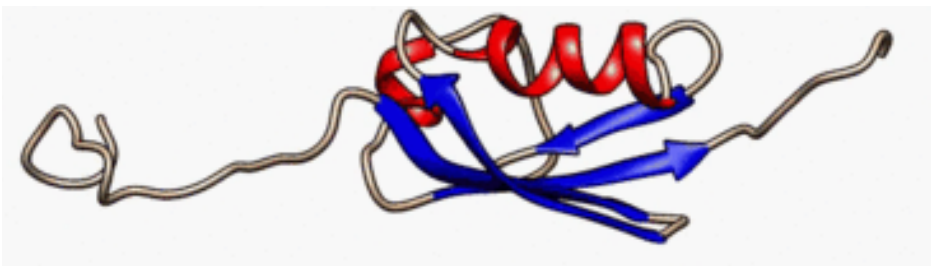
The key: collect and leverage the data

Data modelling pipelines based on alignment and similarity:

- Diversified datasets splitting: training, validation, testing
- A NN classifier as a baseline for the deep learning.

Data augmentation pipelines:

- Label propagation and transformation: A protein -> HMMScan -> domain annotations -> statistical association modelling -> EC/GO labels.
- Mutations in disordered regions have negligible effect, except for mutations to phenylalanine (P), tyrosine (T) and tryptophan (W), which promote order.

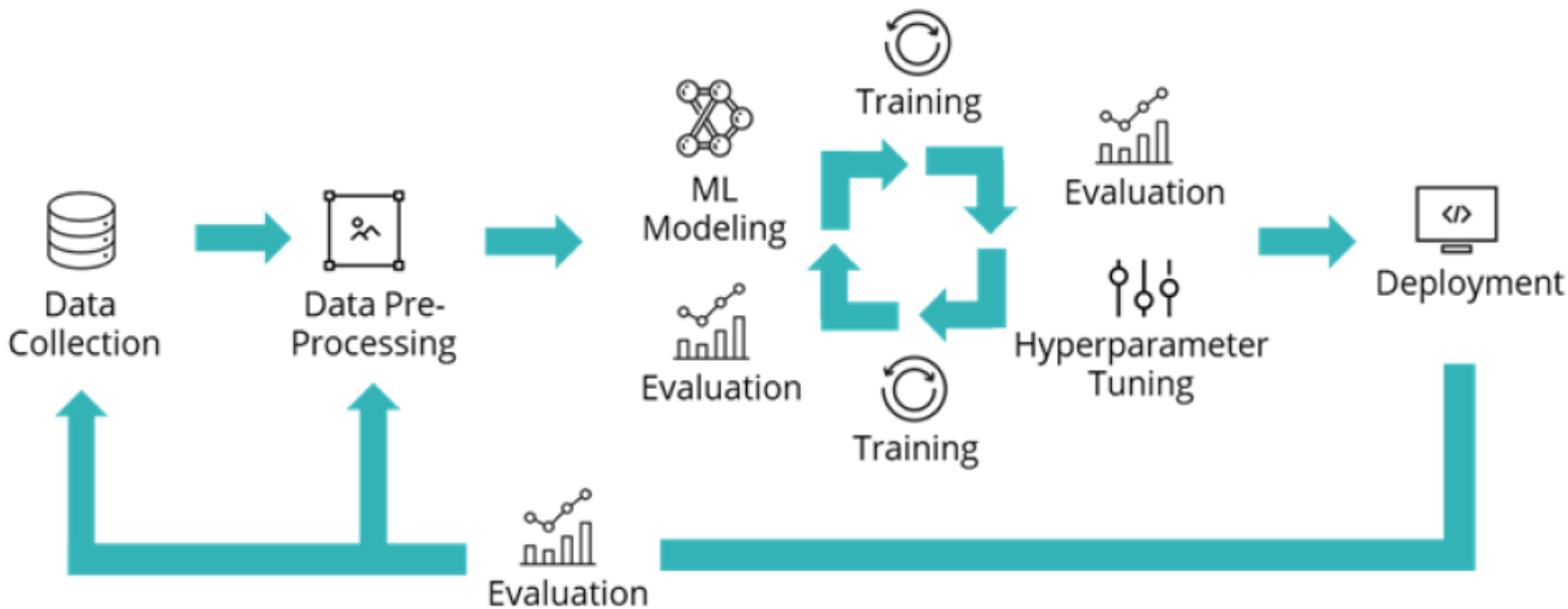


Outline

Industrial R&D for real-world problems

Methods = model-centric AI + data-centric AI

- Implement SOTA solutions & productise them.
- Build modularised AI toolboxes, end-to-end AI service pipelines.



Source: https://dida.do/blog/data-centric-machine-learning

Data collection, curation, and pre-processing take ∼95% of the effort, and are harder to automate.

Outline

Industrial AI Research

Processes and management

1. Suggest research directions and write proposals to the board (CEO and CSO).
2. Plan and Lead research, via task breakdown and an agreed completion timeline tracked by Jira and Confluence.
3. Collaborate, control quality, and maintain conventions using the peer review process, for both code and documentation.

Thanks for your attention.
Questions and discussions are very welcome.

Research topics and interests of Dr. (Amos) Xinshao Wang:

- Deep distance metric learning
- Robust deep learning
- Omics AI + Bioinformatics
- Active learning
- EDA + Data Visualisation

Google Scholar: yOBhB7UAAAAJ
Github: https://github.com/XinshaoAmosWang
Homepage+Blogs: https://xinshaoamoswang.github.io/

References

[1] Bileschi, M. L., Belanger, D., Bryant, D. H., Sanderson, T., Carter, B., Sculley, D., Bateman, A., DePristo, M. A., and Colwell, L. J. Using deep learning to annotate the protein universe. *Nature Biotechnology*, pp. 1–6, 2022.

[2] Cervettini, D., Tang, S., Fried, S. D., Willis, J. C., Funke, L. F., Colwell, L. J., and Chin, J. W. Rapid discovery and evolution of orthogonal aminoacyl-trna synthetase–trna pairs. *Nature biotechnology*, pp. 989–999, 2020.

[3] Elnaggar, A., Heinzinger, M., Dallago, C., Rehawi, G., Yu, W., Jones, L., Gibbs, T., Feher, T., Angerer, C., Steinegger, M., et al. ProtTrans: Toward understanding the language of life through self-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 7112–7112.

[4] Gelman, S., Fahlberg, S. A., Heinzelman, P., Romero, P. A., and Gitter, A. Neural networks to learn protein sequence–function relationships from deep mutational scanning data. *Proceedings of the National Academy of Sciences*, 2021.

[5] Wang, J., Lisanza, S., Juergens, D., Tischer, D., Watson, J. L., Castro, K. M., Ragotte, R., Saragovi, A., Milles, L. F., Baek, M., et al. Scaffolding protein functional sites using deep learning. *Science*, pp. 387–394, 2022.

[6] Wang, X., Kodirov, E., Hua, Y., and Robertson, N. M. Derivative manipulation for general example weighting. *arXiv preprint arXiv:1905.11233*, 2019.

[7] Wang, X., Hua, Y., Kodirov, E., Mukherjee, S. S., Clifton, D. A., and Robertson, N. M. Proselflc: Progressive self label correction towards a low-temperature entropy state. *arXiv preprint arXiv:2207.00118*, 2022.

[8] Wang, X., Hua, Y., Kodirov, E., and Robertson, N. Ranked list loss for deep metric learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 5414–5429, 2022.